Inferring change points and nonlinear trends in multivariate time series: Application to West African monsoon onset timings estimation

Julien Gazeaux, Emmanouil Flaounas, Philippe Naveau, Alexis Hannart

Inferring change points and nonlinear trends in multivariate time series: Application to West African monsoon onset timings estimation

Julien Gazeaux,[1] Emmanouil Flaounas,[1] Philippe Naveau,[2] and Alexis Hannart[3]

Received 7 July 2010; revised 15 November 2010; accepted 9 December 2010; published 1 March 2011.

[1] Time series in statistical climatology are classically represented by additive models. For example, a seasonal part and a linear trend are often included as components of the sum. Less frequently, hidden elements (e.g., to represent the impact of volcanic forcing on temperatures) can be integrated. Depending on the complexity and the interactions among the different components, the statistical inference challenge can quickly become difficult, especially in a multivariate context where the timings and contributions of hidden signals are unknown. In this article we focus on the statistical problem of decomposing multivariate time series that may contain both nonlinear trends and change points (discontinuities), the change points being assumed to occur simultaneously in time for all variables in the multivariate analysis. The motivation for such a study comes from the statistical analysis of the West African monsoon (WAM) phenomenon for which unknown preonset and onset dates occur each year. The impacts of such onsets can be statistically viewed as yearly change points that affect, almost synchronously, trends in observed time series such as daily Outgoing Longwave Radiation and the Intertropical Discontinuity. Our proposed model corresponds to a multivariate additive model with nonlinear trends and possible yearly discontinuities, modeling the onsets. An inference scheme based on a nonlinear Kalman filtering approach is proposed. It enables to identify the different parts hidden in the original multivariate vector. Our inference strategy is tested on simulated data and applied to the analysis of the WAM phenomenon during the period 1979–2008. Our extracted onset dates are then compared to the ones obtained from past studies.

Citation: Gazeaux, J., E. Flaounas, P. Naveau, and A. Hannart (2011), Inferring change points and nonlinear trends in multivariate time series: Application to West African monsoon onset timings estimation, *J. Geophys. Res.*, *116*, D05101, doi:10.1029/2010JD014723.

1. Introduction

[2] Climate time series can often be affected by artificial shifts and/or natural discontinuities due to changes in measurement conditions for the former and physical changes for the latter. To detect and interpret such abrupt and local shifts, many so-called change point statistical procedures have been developed and studied in time series analysis [e.g., *Beaulieu et al.*, 2007]. Current methods simultaneously determine the number of change points and infer their positions. Beyond the specific context of homogenization in climatology [e.g., *Caussinus and Mestre*, 2004], the change point problem is a vast and extensively treated domain of statistics, with applications in econometrics, finance, biology, agronomy and hydrology, among others. A general review of most common approaches can be found in work by *Reeves et al.* [2007]. In a frequentist context, *Davis et al.* [2006] provided a genetic optimization algorithm to extract change points in nonstationary univariate time series using Minimum Description Length principle assuming piecewise autoregressive models. Within a Bayesian framework [e.g., *Chib*, 1998; *Lavielle and LeBarbier*, 2001], *Hannart and Naveau* [2009] recently proposed a fast and efficient algorithm to perform a multiple change point detection technique based in segmenting the time series into subsequences and on prior knowledge derived from past homogenization studies. One common assumption in most change point algorithms is that smooth trends have been removed prior to applying a chosen detection procedure. Basically, this

[1] Laboratoire Atmosphères, Milieux, Observations Spatiales, IPSL, UPMC, UVSQ, CNRS/INSU, Paris, France.
[2] Laboratoire des Sciences du Climat et de l'Environnement, IPSL, CNRS/CEA, Gif-Sur-Yvette, France.
[3] Institut Franco-Argentin d'Études du Climat et ses Impacts, Universitad de Buenos Aires, CNRS/CONICET, Buenos Aires, Argentina.

0148-0227/11/2010JD014723

means that the data under study are assumed to come from a zero mean stationary signal affected by an unobserved change point process that characterizes the timing and the amplitudes of the hidden shifts. For the practitioner, this assumption implies a procedure that has two independent steps: (1) the removal of trends and (2) the extraction of change points. This makes sense in homogenization because meteorologists [e.g., *Caussinus and Mestre*, 2004] classically work with pairwise differences from a set of temperature records, and an artificial shift in one time series should remain in the differences while smooth trends disappear by differencing. In other applications this two-step strategy may not be optimal, and the assumption of a zero mean stationary signal with shifts in step 2 can be challenged. To illustrate this issue, one can imagine two idealized cases. First, two time series, say, of daily methane and ozone recorded at the same station, have a few common artificial discontinuities, e.g., due to changes in the station location. Making the difference between these two series would not necessarily remove trends because methane and ozone may have different low-frequency signatures. The second case could be of two temperature recordings over a climatic homogeneous region in which spatially coherent abrupt changes occur synchronously in time (maybe due to weather regimes modifications or network-wide changes in observing practice). Here having synchronous breakpoints implies that taking the difference between the two time series could greatly diminish the hidden shifts intensity and consequently make it impossible to find change points from this difference. For these two examples one option could be to preprocess each series independently in order to remove the low-frequency components. Subtracting these low-frequency components in order to work with zero mean stationary signals can still be an issue because large change points induce a strong bias in the overall background variance estimation, and consequently, this may lead to estimation errors of these low-frequency components. In addition, any estimation errors produced during the first step (the removal of trends) can propagate other estimation errors into the second step (the change point extraction procedure). Finally, a joint statistical analysis should improve the detection, because the hidden signal is supposed to affect all time series (with different degree). Ideally, it would be of interest to propose a global model and a general inference approach that bypasses the two-step estimation procedure. In its most general form this objective is overly complex because each time series can have its own nonlinear trend and share hidden change points. Consequently, additional assumptions are needed and they should be driven by the application at hand.

[3] The statistical model presented in this study is applied on both simulated and real climatological data. Section 2 provides the background theory on the real data application, which corresponds to the detection of the West African monsoon (WAM) onset and explains the statistical problem concerning the estimation of unknown yearly onsets timings. Section 3 corresponds to the main statistical part of this work. Our statistical model is defined there, and the inference scheme used to estimate unknown quantities is proposed and tested on simulated data. Then this scheme is applied on two variables representative of the WAM onset, for the period 1979–2008. The extracted onsets are compared to past results. Conclusions and perspectives are discussed in section 4. Appendix A provides the technical parts of our algorithm and technical details about our data sets.

2. West African Monsoon Onsets

[4] The West African monsoon (WAM) regulates the rainfall season and is of paramount importance for food security and local economy. The northern WAM propagation interacts with other regional climatic features (such as the African Easterly Waves) which may result in the cyclogenesis budding within the West African coast and eventually the initiation of tropical cyclones [*Thorncroft and Hodges*, 2001].

[5] The rain band associated to the WAM makes part of the seasonal cycle of the Intertropical Convergence Zone (ITCZ). Following the ITCZ intraseasonal cycle, the WAM blows over West Africa from early spring to early autumn, advecting humidity and regulating the overland ITCZ location. The WAM onset corresponds to the abrupt displacement of the ITCZ and the WAM toward the north and signalizes the initialization of the rainy season for the Sahel. To illustrate the relation between the WAM and the ITCZ, Figure 1 presents monthly averages of Outgoing Longwave Radiation (OLR) superimposed over 925 hPa wind circulation patterns for the period 1979–2008. The OLR values are taken from the National Oceanic and Atmospheric Administration (NOAA) archive [*Liebmann and Smith*, 1996] and is used as a proxy for deep convection since low OLR values are associated to the cold cloud tops of convective systems. The OLR data set is interpolated to a 2.5 × 2.5 grid and corresponds to mean daily values. The wind data are taken from the National Center for Environmental Prediction (NCEP) 2 reanalysis [*Kanamitsu et al.*, 2002], also corresponding to mean daily values interpolated to a 2.5 × 2.5 grid. For all months, the ITCZ is marked by low OLR values, and the WAM is represented by the southwest flow. The WAM, due to its charge in humidity, is cooler and more humid than the northeast dry and warm Harmattan wind which originates from the Sahara desert. Hence, a zone with frontal characteristics is created which propagates according to the WAM inland penetration. This frontal zone is referred to as the Intertropical Discontinuity (ITD). Due to the different direction of these two winds, the location of the ITD is determined by the zero isotach of the zonal wind. From May until early June the ITCZ is strong and located over the Guinean coast (approximately at 5N). Similarly, the WAM presents a weak inland intrusion, and hence the ITD is located along 15N (preonset period). On the other hand, from July to August the ITCZ is installed over the Sahel (along 10N), and the ITD reaches 20N (postonset period). The transition from the preonset period to the postonset period is characterized by the significant weakening of convection over the entire region and is detected to occur during late June.

[6] Taking advantage of the zonal symmetry of the ITCZ and the ITD over West Africa, Figure 2 shows Hovmoeller diagrams of three random years (1992, 1998, and 2005) for the OLR values superimposing the ITD location. In these diagrams the OLR values and the ITD location are averaged between 10W and 10E and then smoothed by a moving

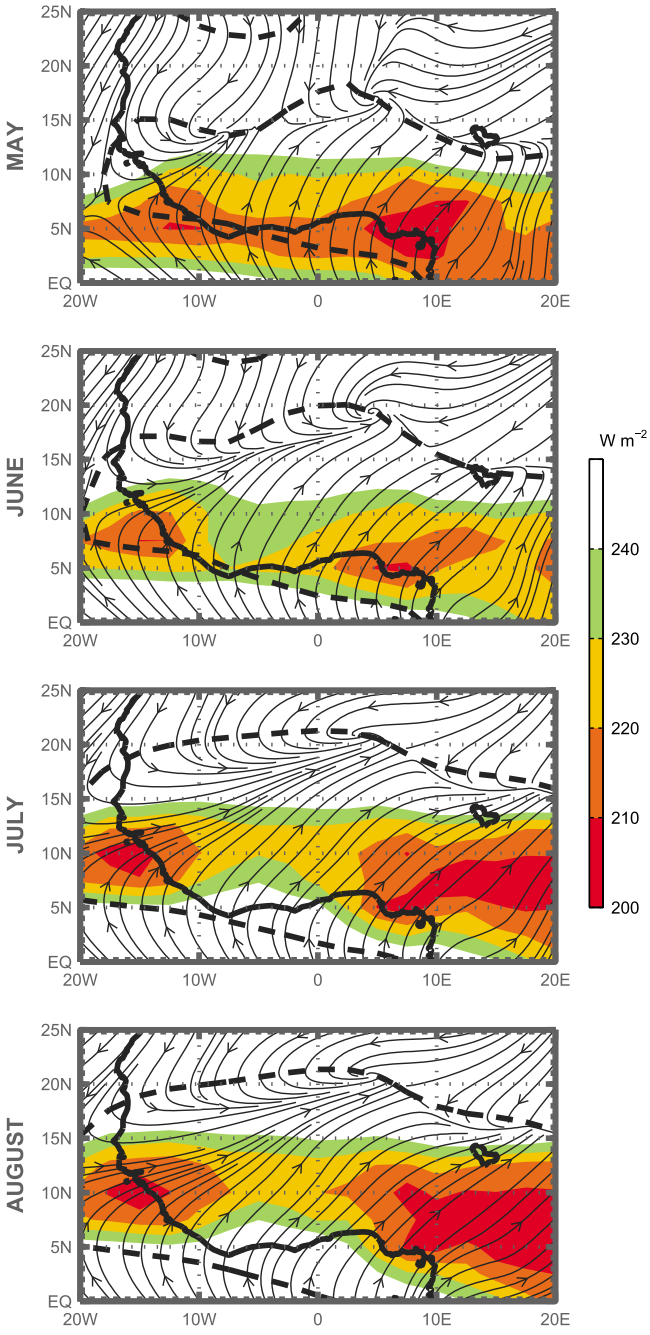

Figure 1. Illustration of the onset phenomena: Monthly averages of the 925 hPa atmospheric circulation and OLR fields from May to August. Thick contour represents the zero zonal wind isotach.

average of 2 days to eliminate intense variability. For the 3 years plotted, it is important to underline the repeat of the same time-latitude pattern for both OLR and ITD. The northward displacement of the ITD is also accompanied by the decrease of OLR values.

[7] Previous studies of rainfall climatology [*Nicholson*, 1981; *Sultan and Janicot*, 2000; *Le Barbé et al.*, 2002] have put into light the intraseasonal cycle of the ITCZ. Detecting the initialization of the rainy period over the Sahel (10N to 20N), which is characterized by abrupt changes in the regional atmospheric circulation and rainfall, is currently an object of active research [*Fontaine et al.*, 2008].

[8] By plotting the average precipitation between 10W and 10E along 15N from the NCEP2 reanalysis database, *Sultan and Janicot* [2003] identified two breaks in the positive rainfall slope, and they interpreted them as pre-onset dates (when the ITD reaches 15N) and onset dates (installation of the ITCZ along 10N). *Fontaine and Louvet* [2006] analyzed rainfall data to define two precipitation indexes. The first one was based on averaging precipitation over the region (10W to 10E and the equator to 7.5N) and the second one over the same longitude band but with different latitudes, from the equator to 20N. Whenever the difference between these two indexes became positive for at least 20 days, an onset was considered to have taken place during the first instant of this period. Finally, *Fontaine et al.* [2008] studied OLR data to determine onset dates by calculating percentages of deep convection occurrences.

[9] Inspired by these different studies, we aim at proposing a unifying statistical approach that can view such onsets as yearly change points that affect, almost synchronously, multivariate time series. Following the aforementioned authors, we construct two time series from two databases. First of all, we construct a time series of daily OLR fields (taken from the NOAA archive) within the Sahel region (10W to 10E and 12.5N to 20N) for each year from 1979 to 2008. The 12.5N boundary of the chosen

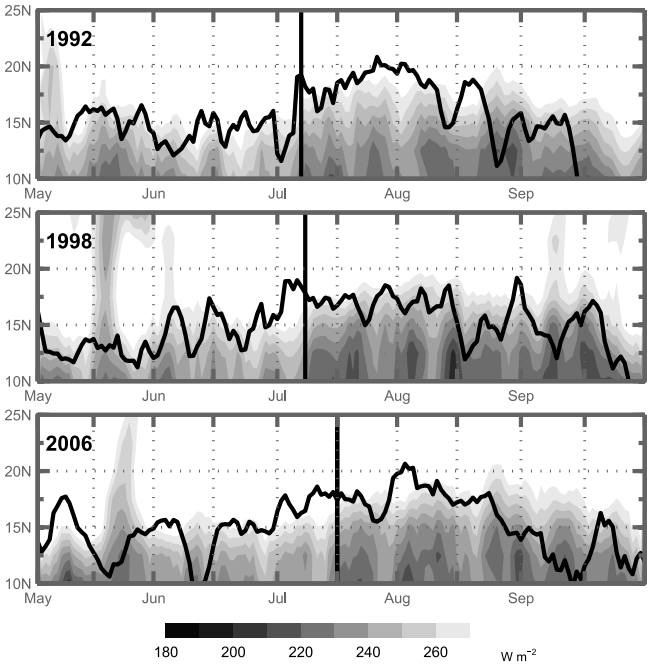

Figure 2. Hovmoeller diagram of OLR. OLR values were averaged from 10W to 10E and smoothed by a moving average of +/−2 days. Thick black line corresponds to the ITD position as the zero zonal wind isotach at 925 hPa. The vertical black bars represent the dates of the onset we estimated. We zoomed the time axis to better show the phenomena.

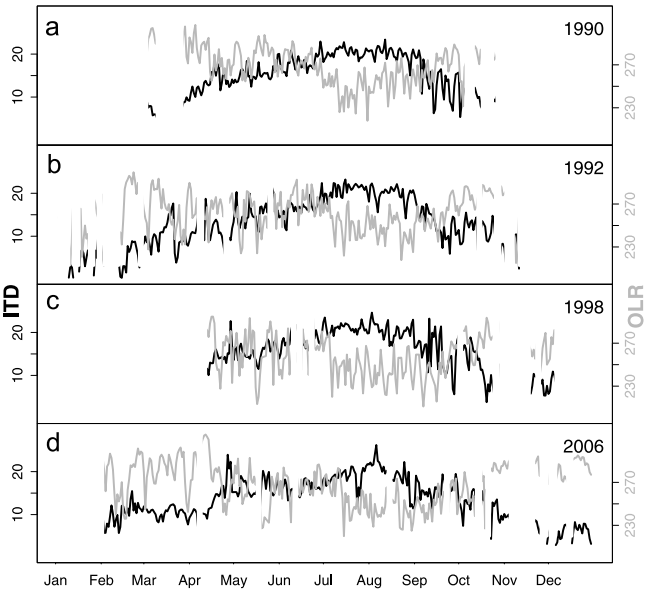

Figure 3. Daily Outgoing Longwave Radiation (OLR) and Intertropical Discontinuity (ITD) time series for four different years: (a) 1990, (b) 1992, (c) 1998, and (d) 2006. The dark and grey lines correspond to ITD and OLR data, respectively. The missing values in ITD are due to the difficulty to calculate the latitude of the zero zonal wind. The ITD unit is latitude, whereas OLR is $W.m^2$.

domain is justified from the fact that is far from the Guinean coast, and it is strongly affected by convection over the Sahel. Hence, convective activity over Sahel before the actual WAM onset would create change point signals which would be detected as spurious onset by our statistical model (as in Figure 12, around 25 May). Second, the NCEP2 reanalysis is used in order to detect the northern reach of the WAM. For this reason we calculate the mean daily location of the ITD, which corresponds to the mean latitude of the zero zonal wind isotach between 10W and 10E.

[10] To illustrate the yearly behavior of such data, Figure 3a displays the daily time evolution of the ITD location (dark line in latitude) and OLR (gray line in $W.m^2$) time series from January 1992 to November 1992. From Figure 3a it is clear that the ITD location steadily increases until the month of August and is followed by a slow decline in Autumn.

[11] The OLR has the opposite low-frequency behavior with a stronger variability. Following the work of *Fontaine and Louvet* [2006] and *Sultan and Janicot* [2003], we postulate that these ITD and OLR signals could contain hidden change points corresponding to the preonset (installation of the ITCZ along 5N) and the onset (installation of the ITCZ along 10N) dates. Hence, besides the high variability observed in Figure 3 and the presence of missing data, the statistical issue at hand in this paper is how can smooth behaviors as well as hidden shifts be estimated by jointly modeling these OLR and ITD time series for the year 1992. The same question can be asked for each year in 1979–2008. To show the year-to-year variability, the years 1990, 1992, 1998, and 2006 are plotted in Figures 3a–3d, respectively. It is an understatement to say change points and trends are not easily identifiable by visual inspection of Figure 3, and nontrivial statistical analysis is needed.

3. Statistical Modeling and Inference

[12] Our statistical model takes its roots in the classical family of state space models. This means that a two-layer structure provides the modeling foundation. The first layer corresponds to the data while the second layer represents the processes of interest which live in the so-called state space. The first and second layers are observed and hidden, respectively. A large body of work on inverse problems, data assimilation, and Bayesian modeling is based on this idea of state space modeling. For example, the well-known Kalman Filter (KF) allows us to estimate the hidden state of a dynamical linear system [e.g., *Kalman*, 1960; *Welch and Bishop*, 1995; *Meinhold and Singpurwalla*, 1983]. The KF has been extended in many ways to take into account nonlinearities and to deal with large data sets. For example, the work of *Evensen* [2006] treats the Ensemble Kalman Filter for data assimilation.

[13] From a methodological point of view, our proposed statistical model stems from the work of *Guo et al*. [1998], who studied an extracting procedure, not for change points but for pulse-like signals in univariate hormone time series. The generic shape of the hidden signal given by *Guo et al*. [1998] corresponded to a peak followed by a sharp decrease, while a stepwise function is the object of interest in most change point procedures. J. Gazeaux et al. (Extracting common pulse-like signals from multiple ice core time series, submitted to *Computational Statistics and Data Analysis*, 2011) improved Guo's approach by extending it from the univariate case to the multivariate case and by applying it to the problem of volcanic forcing extraction from multivariate proxy data. Now, by building on the multivariate approach studied by J. Gazeaux et al. (submitted manuscript, 2011), we propose to capture change points and smooth trends. Due to the altered nature of the extracted signal and the model constraints imposed by our monsoon application, this extension is far from trivial. A new model is needed, and the statistical inference procedure has to be modified. Our proposed model corresponds to a multivariate additive model with nonlinear trends and possible yearly discontinuities, the latter captured yearly onsets. While an autoregressive cubic spline representation is used to depict different smooth trends, another autoregressive model with noncontinuous innovations mimics the change points dynamic. Blending together these two autoregressive models offers a modeling flexibility and removes some classical hypotheses; no linear assumption is required. To balance this flexibility in the low-frequency part of the spectrum, we impose that the unknown breakpoints occur synchronously in time in all variables; see the WAM onsets application.

[14] Concerning our notations, $y_j(t)$ represents the value of the jth variable of interest for day t. For example, $y_1(t)$ and $y_2(t)$ could correspond to the daily OLR and ITD values in 1990 (see Figure 3a). Such random variables are assumed to come from the following additive model:

$$y_j(t) = f_j(t) + \beta_j x_t + \epsilon_j(t) \quad \text{with} \quad j = 1, \ldots, J \text{ and } t = 1, \ldots, T, \tag{1}$$

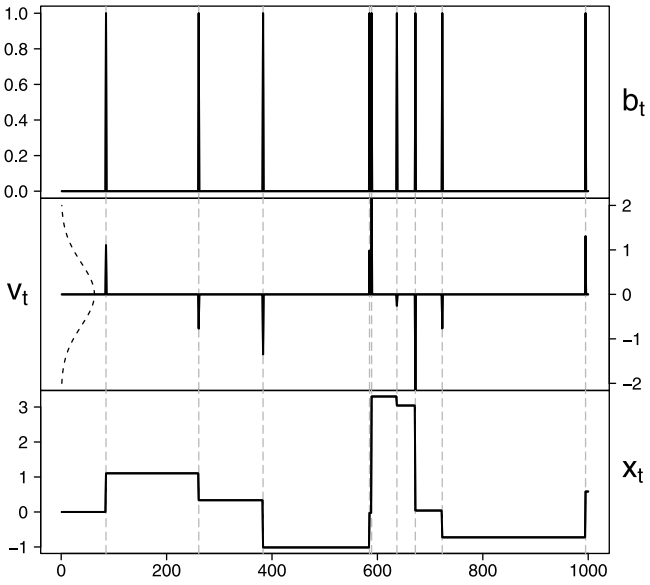

Figure 4. Random realizations from equations (2) and (3). The top, middle, and bottom panels show the Bernoulli signal b_t, the hidden impulse v_t in (3), and the hidden stepwise x_t obtained from (2), respectively.

where $f_j(t)$ represents the smooth trend specific to the jth time series, x_t represents the change points signal common to all time series, β_j is the scaling factor of the x_t impact to the jth time series, and finally, $\epsilon_j(t)$ is a zero mean independent and identically distributed (iid) Gaussian noise with variance σ_j^2. The elements of the sum (1) are assumed to be mutually independent. Equation (1) clearly indicates that each time series can have a different trend with its own noise and a common element x_t whose impact is modulated by β_j. A strong assumption of the method is, through β_j of equation (1), the proportionality of the break points occurring at the same time. If all β_j have the same sign, the breaks have the same effect, either "positive" or "negative"; on the other hand, if the β_j have opposite signs, the breaks will have opposite effects: one will be "positive" while the other will be "negative" and vice versa. We suppose in (1) that there is not a missing value, i.e., with the constant sampling rate $t = 1,\ldots,T$. But Figure 3 shows missing values. Our model and our inference can handle this case by transforming the time axis $1,\ldots,T$ into $t_1,t_2,\ldots,t_k$. For sake of clarity, we still prefer to present our method with $t = 1,\ldots,T$. As the hidden signal x_t should capture the onsets dynamic, we follow the classical view of modeling change points as a random stepwise function. Here this stepwise behavior is represented by an autoregressive model of order one

$$x_t = x_{t-1} + v_t, \tag{2}$$

where the random variable v_t either equals zero or a zero mean random Gaussian vector z_t with variance σ_v^2, i.e.,

$$v_t = \begin{cases} 0 & \text{if } b_t = 0 \text{ with probability } 1-\pi, \\ z_t & \text{if } b_t = 1 \text{ with probability } \pi, \end{cases} \tag{3}$$

with $x(0)$ set to zero; b_t is a Bernoulli iid process, either equal to one or zero with probability π and $1-\pi$, respectively. The process b_t drives the occurrences of the impulses. The Gaussian variables z_t are iid and independent of b_t.

[15] To understand equation (3), we refer to Figure 4. Figure 4, bottom, shows one random realization of the stepwise behavior of x_t defined (2). The elements of this autoregressive process, i.e., b_t and v_t, are displayed in Figure 4, top and middle. Although autoregressive, the process x_t is zero mean but not stationary because its variance increases linearly with time, $\mathbb{V}\mathrm{ar}(x_t) = \pi t\sigma_v^2$. In our WAM application, this is not a fundamental issue; because the yearly probability of observing a change point π is very small, we expect to have one or two change points (preonset and onset) per year. This implies that the yearly largest $\mathbb{V}\mathrm{ar}(x_t)$ should be about $2\sigma_v^2$ and does not explode with time. This also justifies that we analyze our data year-per-year and not the entire period 1979–2008 in one run (this is compounded with the fact that yearly trends have a strong year-to-year variability; see Figure 3). Hence the hidden stepwise x_t obtained from equation (2) is unlikely to produce its own trend. This implies that only the component $f_j(t)$ in equation (1) should capture the low frequency in $y_j(t)$.

[16] To model the trend $f_j(t)$, we opt for a cubic smoothing spline representation [*Wahba*, 1978]. The latter can be described as a multivariate autoregressive model of order one [*Wecker and Ansley*, 1983]

$$\mathbf{F}_j(t) = B\mathbf{F}_j(t-1) + \mathbf{E}_j(t) \tag{4}$$

where $\mathbf{F}_j(t) = \begin{bmatrix} f_j(t) \\ f_j'(t) \end{bmatrix}$ represents a bivariate vector that includes (f_j) and its first derivative (f_j'), the matrix B equals $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, and $\mathbf{E}_j = \begin{bmatrix} E_{f_j} \\ E_{f_j'} \end{bmatrix}$ is a two-dimension zero mean Gaussian vector with covariance matrix equal to $\lambda_j {\sigma_{f_j}}^2/(\mathrm{j} + \mathrm{k} - 1)(\mathrm{j} - 1)!(\mathrm{k} - 1)!$ where λ_j represents the smoothing parameter, and σ_{fj} a positive constant. Equation (4) implicitly means that the trend f_j and its first derivative f_j' are assumed to be continuous.

[17] Choosing equations (2) and (4) to model the unknown trends and the hidden change points dynamic brings an important inferential benefit because our model can be rewritten as a classical linear state space model of the form:

$$\begin{aligned} Y_t &= HX_t + E_t, \\ X_t &= \Phi X_{t-1} + E_t^*, \end{aligned} \tag{5}$$

where the first equality corresponds to the so-called observation equation, and the second one corresponds to the so-called state equation [e.g., *Meinhold and Singpurwalla*, 1983]. To clarify the link between equations (5) and (2)–(4), we write below the form of the elements of equation (5) for $J = 2$, i.e., the case of the daily OLR and ITD random variables, in function of the components of equations (2)–(4)

$$\begin{aligned} Y_t &= [y_1(t), y_2(t)]^T,\ X_t = [v_t, x_t, \mathbf{F}_1, \mathbf{F}_2]^T \\ &\text{with } \mathbf{F}_j(t) = \left[f_j(t), f_j'(t)\right]^T \end{aligned} \tag{6}$$

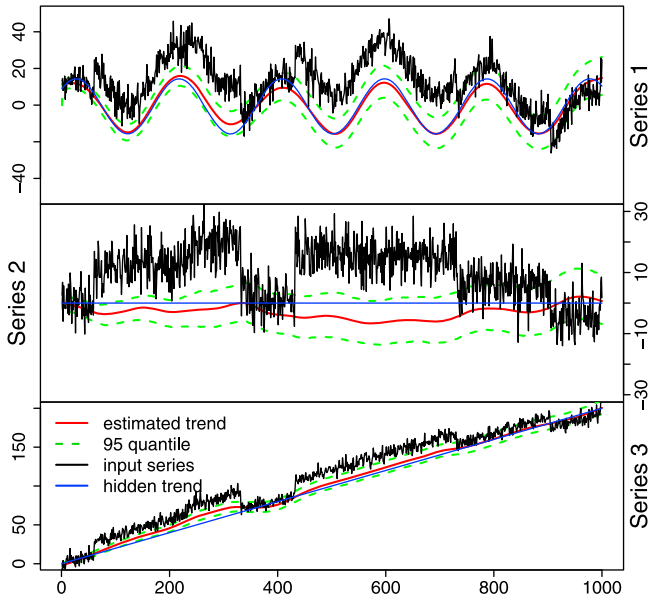

Figure 5. Extraction obtained from three simulated time series. The blue and red lines correspond to the true and estimated trends, respectively. The 95% confidence interval is represented by the green dotted lines.

and

$$H = \begin{bmatrix} 0 & \beta_1 & 1 & 0 & 0 & 0 \\ 0 & \beta_2 & 0 & 0 & 1 & 0 \end{bmatrix} \quad E_t = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}, \Phi = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & B & 0 \\ 0 & 0 & 0 & B \end{bmatrix},$$

$$E_t^* = \left[0, 0, E_{f_1}, E_{f_1'}, E_{f_2}, E_{f_2'}\right]^T$$

$$\text{and } \mathbb{C}ov(\mathbf{E}_j) = \lambda_j \sigma_j^2 \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{bmatrix}.$$

The advantage of transforming equations (2)–(4) into the state space form equation (5) is that developments about KF can serve as building blocks for our inference procedure. By construction, the observation noise E_t and the state equation noise E_t^* are uncorrelated.

[18] As in any KF algorithm, our main goal is to estimate the hidden state X_t given current and past observations, i.e., given the vector $Y_{1:t} = (Y_1,\ldots,Y_t)^T$. We cannot directly apply the classical KF to reach this aim because the binary vector b_t in the definition of v_t makes the hidden vector X_t non-Gaussian; see Figure 5. Still, as proposed by *Guo et al.* [1998], two ideas can be followed to remove this inference block. First, by conditioning on the value of b_t, either one or zero, we can sequentially compute the conditional expectation and variance of the two random variables $[X_t|Y_{1:t},\ b_t = 1]$ and $[X_t|Y_{1:t},\ b_t = 0]$ at time t, whenever those mean and variance are available at time $t-1$. The random variable $[X_t|Y_{1:t},\ b_t = 1]$ corresponds to the hidden state given observations up to time t and having observed a change point at time t and $[X_t|Y_{1:t},\ b_t = 0]$ is the same except that no change point has been observed at time t.

[19] The second idea is to approximate the non-Gaussian distribution of $[X_t|Y_{1:t}]$, because of b_t, by a Gaussian one whose first and second moments equal those of $[X_t|Y_{1:t}]$. This approximation that works well in practice (see section 4) allows us to update the mean and variance of the variable of interest $[X_t|Y_{1:t}]$ as follows (with obvious notations described in Appendix A)

$$\begin{aligned} \hat{X}(t|Y_{1:t}) &= q_t^0 \hat{X}(t|Y_{1:t}, b_t = 0) + q_t^1 \hat{X}(t|Y_{1:t}, b_t = 1) \\ \hat{\Sigma}(t|Y_{1:t}) &= q_t^0 \hat{\Sigma}(t|Y_{1:t}, b_t = 0) + q_t^1 \hat{\Sigma}(t|Y_{1:t}, b_t = 1) \\ &\quad + \sum_{i=0}^{1} q_t^i \left(\hat{X}(t|Y_{1:t}, b_t = i) - \hat{X}(t|Y_{1:t})\right)^2, \end{aligned} \tag{7}$$

where q_t^1 (resp.q_t^0) is the occurrence probability of having (not having, respectively) a breakpoint at time t given $Y_{1:t}$, and it equals (via Bayes' theorem)

$$\begin{aligned} q_t^0 &\doteq Pr(b_t = 0|Y_{1:t}) = \frac{1-\pi}{Pr(Y_t|Y_{1:t-1})} Pr(Y_t|Y_{1:t-1}, b_t = 0) \\ q_t^1 &\doteq Pr(b_t = 1|Y_{1:t}) = \frac{\pi}{Pr(Y_t|Y_{1:t-1})} Pr(Y_t|Y_{1:t-1}, b_t = 1) \end{aligned} \tag{8}$$

where $\forall i = 0,1, Pr(Y_t|Y_{1:t-1})$ and $Pr(Y_t|Y_{1:t-1}, b_t = i)$ represent the conditional density of the random variables $[Y_t|Y_{1:t-1}]$ and $[Y_t|Y_{1:t-1}, b_t = i]$, respectively. As for the classical KF, the conditional mean and variance $\hat{X}(t|Y_{1:t}, b_t = i)$ and $\hat{\Sigma}(t|Y_{1:t}, b_t = i)$ can be expressed in terms of previous expressions obtained at time $t-1$. See Appendix A for more details. The estimation of the state vector at every time $t = 1,..T$ regarding the available observation $Y_{1:T}$ is obtained via the Fixed Interval Smoother, which is

$$\begin{aligned} \hat{X}(t|Y_{1:T}) &= \hat{X}(t|Y_{1:t}) + C_t\left[\hat{X}(t+1|Y_{1:T}) - \hat{X}(t+1|Y_{1:t})\right] \\ \hat{\Sigma}(t|Y_{1:T}) &= \hat{\Sigma}(t|Y_{1:t}) + C_t\left[\hat{\Sigma}(t+1|Y_{1:T}) - \hat{\Sigma}(t+1|Y_{1:t})\right] C_t' \end{aligned} \tag{9}$$

where

$$C_t = \hat{\Sigma}(t|Y_{1:t}) \Phi \hat{\Sigma}(t+1|Y_{1:t})^{-1} \tag{10}$$

For more details about these calculations, see Appendix A. So far, we have assumed that the parameters (β_j, π, σ_v, σ_{fj}) were known. This is not true in practice. They are derived through an iterative maximum likelihood estimation computed after a rough estimation of the trend of each time series. This method of approach is successfully used by *Guo et al.* [1998].

[20] To assess the quality of our algorithm, we apply it to simulated trivariate time series defined as follows. The first, second and third smooth trends are equal to $f_1(t) = 10 + 15\sin(\pi(t+20)/90)$, $f_2(t) = 0$ and $f_3(t) = 2t$, respectively. In Figure 5, we can observe the three hidden trends (blue solid lines) and three random realizations (black lines) affected by zero mean Gaussian noises with variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 5$ and a random change point process with $\pi = 0.01$ and $\sigma_v^2 = 1.0$. The scaling coefficients are chosen such that $(\beta_1, \beta_2, \beta_3)^T = (20, 15, 20)^T$. Additional tests (available under request) show that our multivariate method represents an improvement over its univariate counterpart, i.e., applying independently our model to each individual time series.

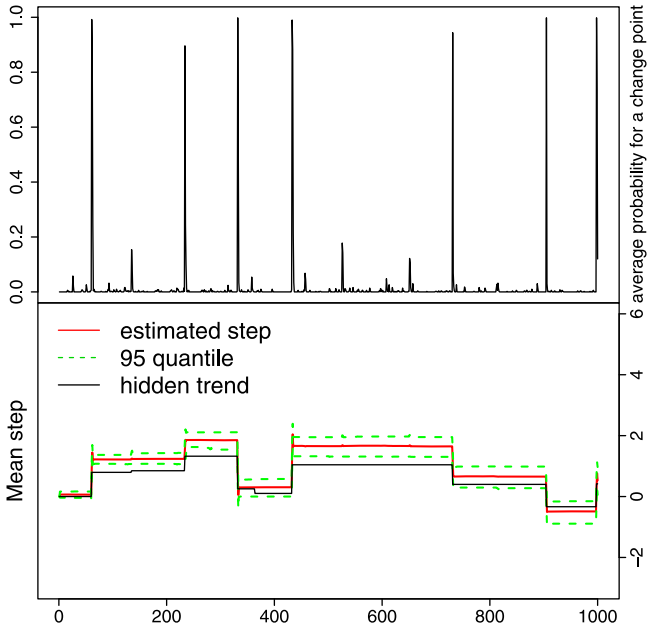

Figure 6. (top) The estimated probability of observing change points simultaneously in the three time series displayed in Figure 5. (bottom) Comparison of the true (black) x_t defined by (2) and its estimate (red) with their 95% confidence interval (dotted green lines).

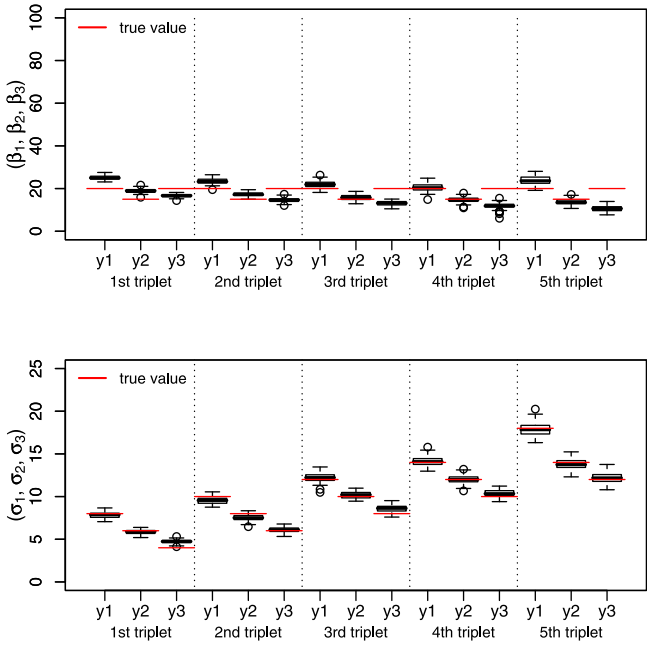

Figure 7. Box plots from 500 simulations with $\beta^T = (20, 15, 20)^T$ and $\pi = 0.01$ and the trends of Figure 5. The x axis corresponds to five different combinations of the triplet $(\sigma_1, \sigma_2, \sigma_3)^T = (8, 6, 4)^T, (10, 8, 6)^T, (12, 10, 8)^T, (14, 12, 10)^T$ or $(18, 14, 12)^T$. Under these five sets of noise levels, the top panel compares the true trivariate β (red horizontal lines) with the box plot of its estimate, and the bottom panel displays the same result but for $(\sigma_1, \sigma_2, \sigma_3)$.

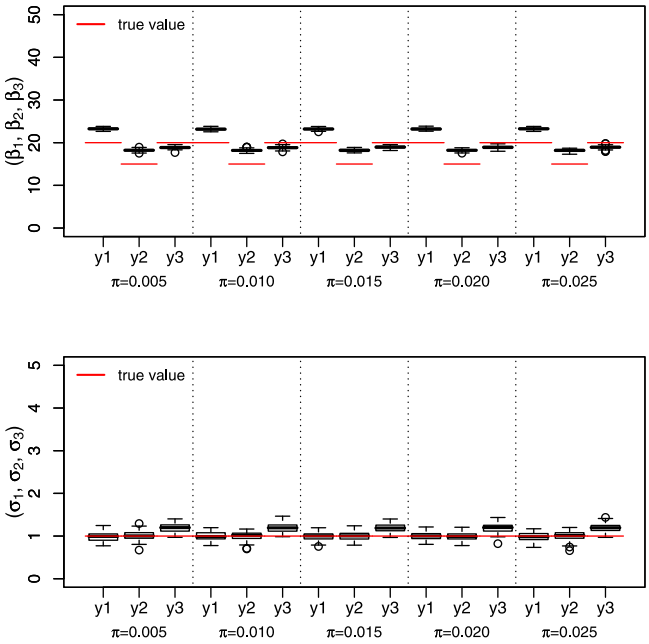

Figure 8. Same as Figure 7 but with a fixed $(\sigma_1; \sigma_2; \sigma_3)^T = (1.0; 1.0; 1.0)$ and five different $\pi = 0.005$, 0.010, 0.015, 0.020, or 0.025.

[21] The red lines correspond to the estimated trends with their 5 and 95 quantiles shown in dotted green lines.

[22] In Figure 6, the three black lines represent the input of the model, whereas the different colored lines are the output of the extraction, i.e., the estimated parts of X_t. Figure 6, top, displays the estimated probabilities of observing change points, and Figure 6, bottom, compares the true x_t and its estimate. Graphically, the timing and amplitudes of the change point appear to be well-estimated. Only the smallest shifts at approximately $t = 160$ and $t = 380$ are associated with low probabilities of about 0.2 and less than 0.1.

[23] The single simulation analysis shown in Figures 5 and 6 is obviously not sufficient to conclude about the overall performance. Rather it has to be understood as a graphical example of the possible outputs available from our approach. To improve our understanding of the limits and advantages of our method, we apply our algorithm to different sets of parameters. For each set, 500 simulations are randomly generated, and box plots of the parameters of interest are plotted. For example, when fixing $\beta = (20, 15, 20)^T$ and $\pi = 0.01$, the x axis of Figure 7 corresponds to five different combinations of the triplet $(\sigma_1, \sigma_2, \sigma_3) = (8, 6, 4)^T, (10, 8, 6)^T, (12, 10, 8)^T, (14, 12, 10)^T$ or $(18, 14, 12)^T$. Under these five sets of noise levels, Figure 7, top, compares the true trivariate β (red horizontal lines) with the box plot of its estimate, and Figure 7, bottom, displays the same result but for $(\sigma_1, \sigma_2, \sigma_3)$. Overall, the noise variances are well-estimated while the β's have a slight bias when the latter is large. In addition, the noise level does not greatly affect the quality of our estimation.

[24] Figure 8 is the same as Figure 7 but with a fixed $(\sigma_1, \sigma_2, \sigma_3) = (1.0, 1.0, 1.0)^T$ and five different $\pi = 0.005$,

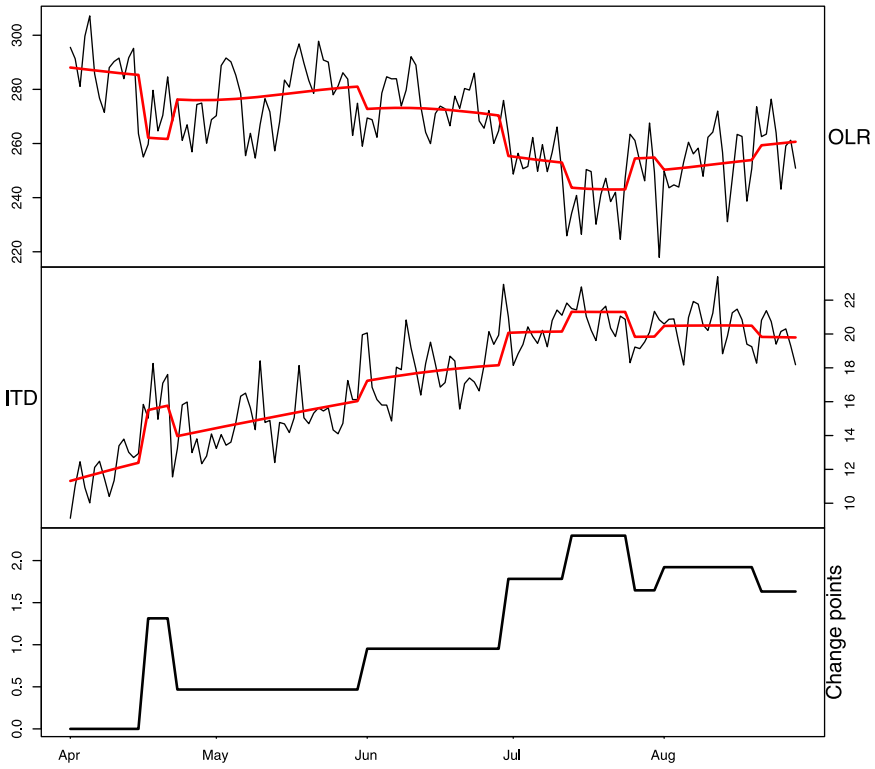

Figure 9. Statistical treatment of the 1990 OLR and ITD time series from Figure 3a. The red line corresponds to the estimated trend $f_1(t)$ and $f_2(t)$ from equation (1). The bottom panel displays the extracted hidden change point signal x_t from equation (2).

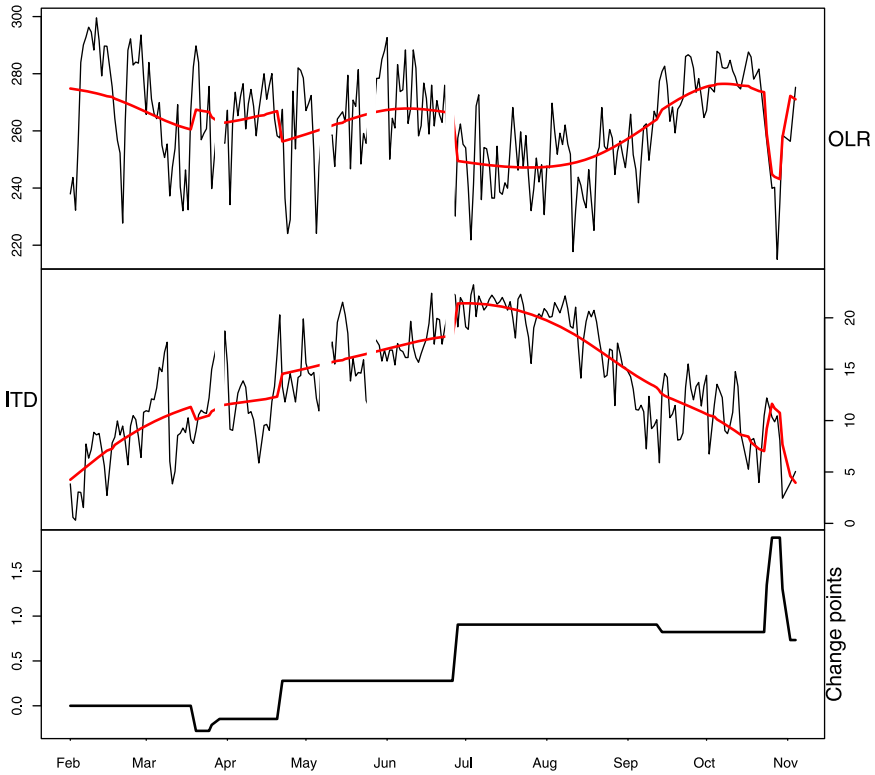

Figure 10. Statistical treatment of the 1992 OLR and ITD time series from Figure 3b. The red line corresponds to the estimated trend $f_1(t)$ and $f_2(t)$ from equation (1). The bottom panel displays the extracted hidden change point signal x_t from equation (2).

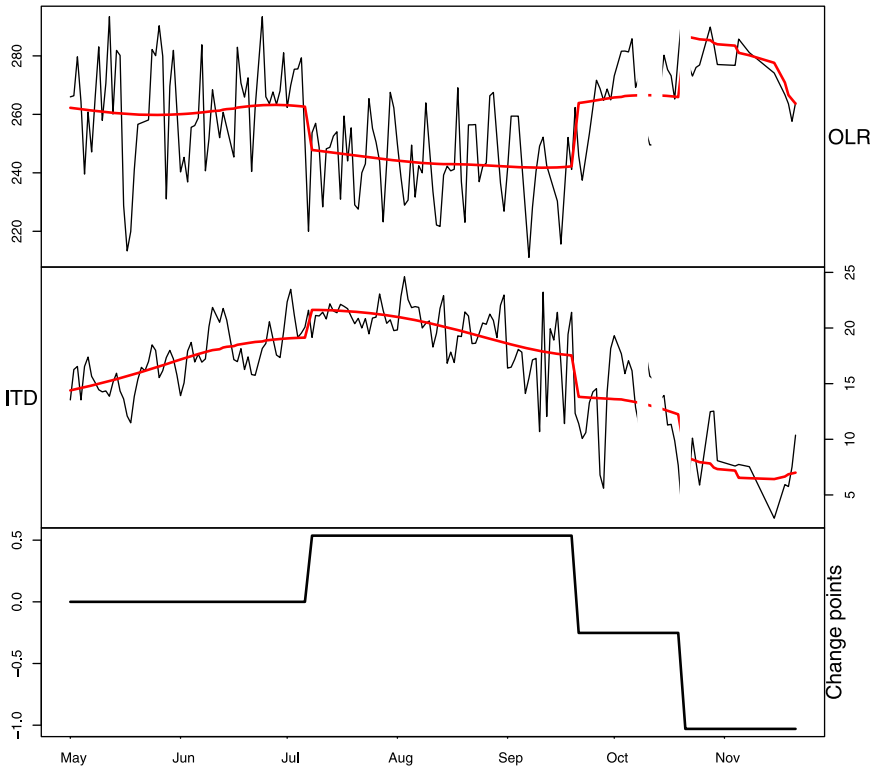

Figure 11. Statistical treatment of the 1998 OLR and ITD time series from Figure 3c. The red line corresponds to the estimated trend $f_1(t)$ and $f_2(t)$ from equation (1). The bottom panel displays the extracted hidden change point signal x_t from equation (2).

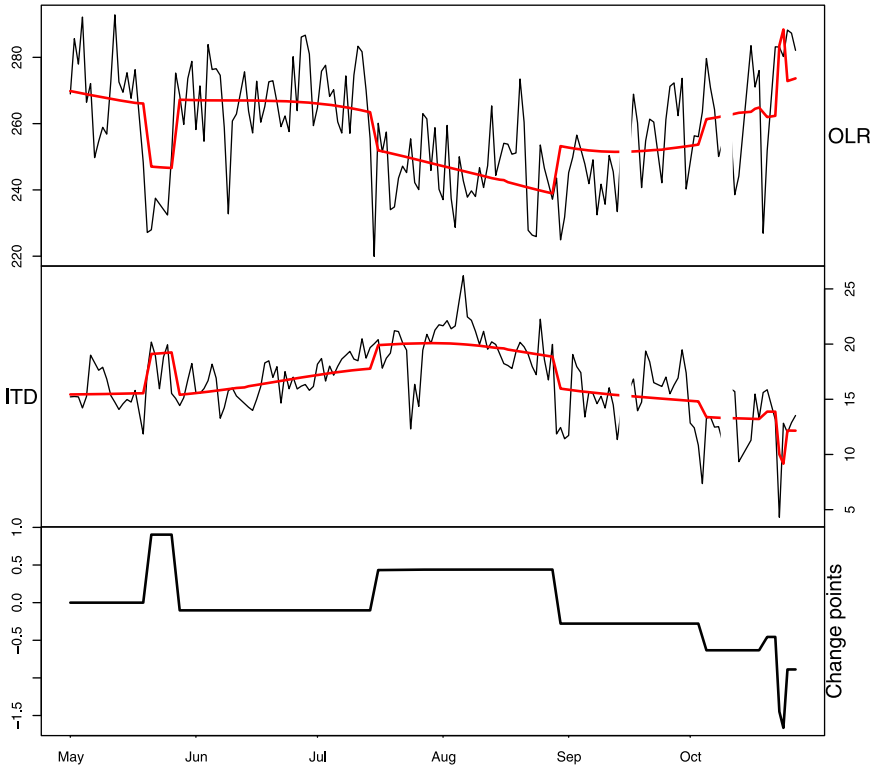

Figure 12. Statistical treatment of the 2006 OLR and ITD time series from Figure 3d. The red line corresponds to the estimated trend $f_1(t)$ and $f_2(t)$ from equation (1). The bottom panel displays the extracted hidden change point signal x_t from equation (2).

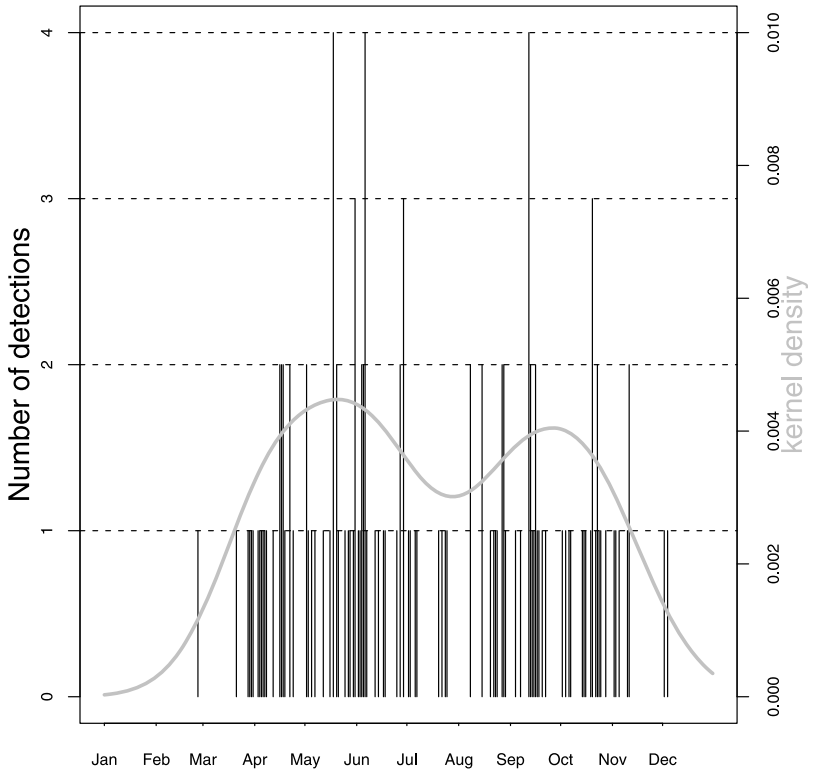

Figure 13. The whole detected change points of each year of WAM time series from 1979 to 2008 with $q_t^1 > 0.5$. The smooth lines represent the density probability calculated with a Gaussian kernel as explained by *Parzen* [1962].

0.010, 0.015, 0.020, or 0.025. These graphs show that changing the number of change points, i.e., driven by π, does not have a strong effect on the estimation of the β's and of the σ_i's.

4. WAM Results and Discussion

[25] Our statistical model and inference method are now applied to the bivariate vector composed of the OLR and ITD time series described in section 2 for each year starting in 1979 and ending 2008. To interpret these outputs, we focus our attention on the 4 years (1990, 1992, 1998, and 2006) introduced in Figure 3. The top and middle panels of Figures 9, 10, 11, and 12 show, in red, the addition of the estimated trends ($f_j(t)$) and the extracted break signals ($\beta_j x_t$) of equation (1) for the OLR and the ITD, respectively. A visual inspection tends to indicate that the trends are well-estimated. Note that the trends ($f_j(t)$) do not have a physical meaning here. Nevertheless, their estimations are necessary to successfully detect onset dates, as the estimations of both signals ($f_j(t)$) and x_t of equation (1) have to be calculated simultaneously. Indeed, statistical methods generally require stationary time series to be reliable. The calculation of $f_j(t)$ using an autoregressive spline estimation (e.g., equation (4)) can be considered as the stationarization process of our method. In other applications, as for instance homogenization problems (i.e., the detection of artificial shifts in time series [e.g., *Caussinus and Mestre*, 2004]), these trends would have physical meaning. Concerning the change points estimation, the extracted common signals seem to be reasonable. For example, Figure 10 clearly indicates an onset around the end of June 1992, and spurious change points appear in the spring and October 1992. Those latter changes are due to poor data quality (missing data, edge effects) and should be disregarded as obvious artifacts in the context of WAM onsets. The same type of reasoning can be employed for the years 1990, 1998, and 2006. Both a pre-onset (in June) and an onset (in July) can be easily identified for years as in Figure 9, while this is not possible for others; see 1992.

[26] For some years like 1988 and 1991, we do not detect any significant onset because we do not force our model to find a specific number of change points. We believe that is a strength, "the data speak for themselves," without a strong a priori on the yearly change point number, and consequently, if the time series are too noisy or the onset is too weak, then there is no detection.

[27] The whole detected break points are illustrated in Figure 13. Each histogram bar represents the number of detected change points per day along the year with probability of occurrence $q_t^1 > 0.5$. Figure 13 displays a bimodal density (black smooth line calculated as a kernel density [cf. *Parzen*, 1962]). The first mode corresponds to both onset and preonset signal mixing together around June, and the second mode corresponds to the end of the monsoon season.

[28] We are more particularly interested in the first mode and discriminate preonset from onset signal thanks to knowledge from previous work on WAM [i.e., *Sultan and Janicot*, 2003]. Some detected change points (as the one occurring in April 1992; see Figure 10) can be considered as

Table 1. Comparison Between Our Detected Onset Dates and the Ones of *Fontaine et al.* [2008]

	1979	1980	1981	1982	1983	1984	1985	1986	1987	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Gazeaux et al.	10 Jul	26 Jun	25 Jun	5 Jun	-	10 Jul	-	6 Jun	9 Jul	16 Jun	29 Jun	-	6 Jul	-	-	21 Jun	14 Jun	25 Jul	7 Jul	28 Jun	3 Jul	16 Jun	15 Jul	10 Jul	-	10 Jun	15 Jul	2 Jul	23 Jun
Fontaine 08(1)	17 Jun	16 Jun	2 Jul	22 Jun	22 Jun	11 Jul	22 Jun	22 Jun	27 Jun	1 Jul	26 Jun	11 Jul	20 Jun	21 Jun	21 Jun	16 Jun	12 Jun	6 Jun	1 Jul	11 Jul	25 Jun	1 Jul	16 Jun	16 Jun	5 Jul	-	-	-	-
Fontaine 08(2)	22 Jul	1 Jul	12 Jul	1 Jul	2 Jul	5 Aug	17 Jun	7 Jul	2 Jul	22 Jul	11 Jul	5 Aug	25 Jun	26 Jun	1 Jul	16 Jun	12 Jun	6 Jun	1 Jul	1 Sep	5 Jul	11 Jul	5 Aug	26 Jun	15 Jun	-	-	-	-

spurious change points (i.e., not an onset signal) [see *Sultan and Janicot*, 2003].

[29] Table 1 compares our results with the two different estimated dates by *Fontaine et al.* [2008]. Although not directly comparable because they were not derived from the same data, most of our dates fall between the proposed dates from *Fontaine et al.* [2008] or differ for about a week.

[30] Figure 14 shows the frequency of our estimated WAM preonset and onset dates for the period 1979–2008. The former dates occur around the beginning of June, and the latter occur around the beginning of July. The onset dates occurring on average on 30 June (with a standard deviation equal to 10 days) are consistent with the climatological date of 24 June found by *Sultan and Janicot* [2003]. Those authors have detected the central date of the transitional period, in contrast to our analysis which reveals the beginning of the postonset period. The preonset dates as determined by the authors (14 May) do not seem to be accurate with the average date of 2 June (with a standard deviation equal to 8 days) found in this study. The reasons for this inconsistency should be highlighted. In the work of *Sultan and Janicot* [2003] the preonset date is determined only by the ITD location. Our results on the preonset date show difficulties to capture this event, i.e., an abrupt northward propagation of the WAM before its onset. The preonset dates are mostly associated to a rather "anomalous" climatological cycle of deep convection over West Africa. For all years the preonset period is well-defined presenting the ITCZ over the Guinean coast. However, for some years the WAM onset comes after no or even two transitional periods with intermediate phases being embedded. These intermediate phases are characterized by the presence of convective activity over both the Guinean coast and the Sahel. This might create confusion on the WAM onset detection. If no abrupt signal is detected (due to a smoothed ITCZ cycle) by the model, then the WAM onset may not be defined (e.g., in 1993), or if intermediate phases are present, then the model may detect more monsoon jumps (e.g., in 2002). In that case, two dates are defined: a "preonset" and an "onset" date. These intermediate phases may be the reason for the large difference of the onset dates for some years, as in 2002, between this study and the study of *Fontaine et al.* [2008].

[31] Finally, to conclude this paper, we would like to say a few words of caution. The detection of the onset appears to be complex. Under the definition of the onset by *Sultan and Janicot* [2003], we expected to find a single clear change point occurring around the end of June, but our methodology clearly detects more than one change point per year. The differences open new questions. Is there really a unique date for the yearly onset? Are our OLR and ITD data the most appropriate time series for detections?

[32] Convection over West Africa is a result of complex dynamics and different forcings from regional or larger-scale climate. Hence, the evolution of the localization of convection could be considered somewhat independent of the ITD. Our statistical approach on the WAM onset could be optimized by using other elements of the West African climate associated to convection and thus eliminating any "false onset". This issue is a perspective for future studies.

[33] To summarize this article, we recall to the reader that our objective was to propose a nonlinear statistical extrac-

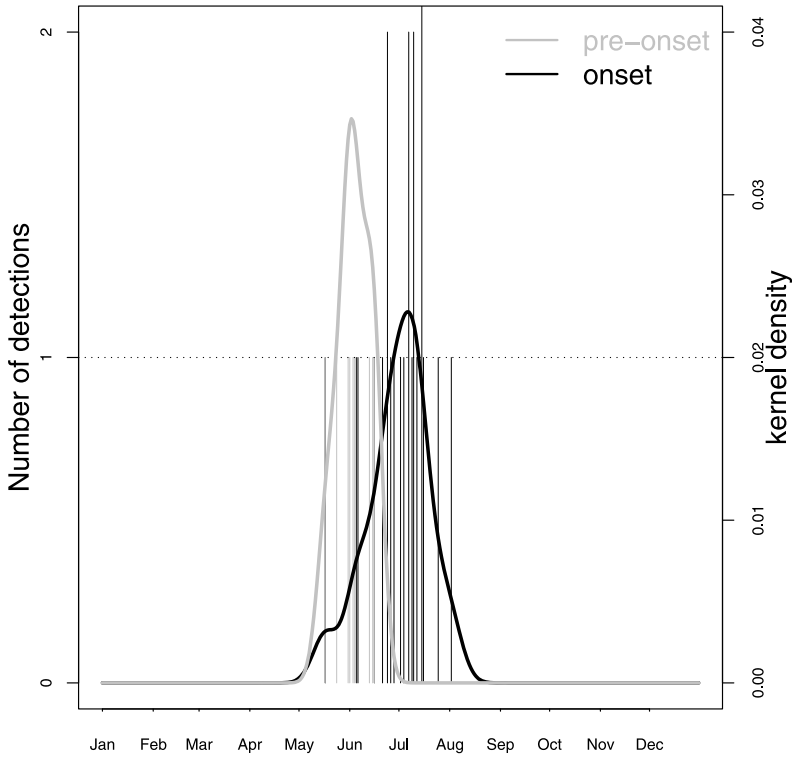

Figure 14. The frequency of our estimated WAM preonset and onset dates for the period 1979–2008. The grey and black colors correspond to preonset dates occurring around the beginning of June and onset dates around the beginning of July, respectively. The smooth lines represent the density probability calculated with a Gaussian kernel as explained by *Parzen* [1962].

tion method that can both infer individual smooth trends and common change points in multivariate time series. From the simulation study, it appears that the proposed inference procedure based on Kalman filtering ideas works adequately. We illustrate the applicability of our method by detecting preonset and onset dates of convection over the Sahel related to the WAM. For this specific application, the advantage of our approach resides in the global representation of uncertainties, and it does not contradict similar studies based on different data and simpler statistical techniques. The estimation of the onset dates distribution of Figure 14 could also likely be treated as relevant a priori information for future prediction studies.

[34] The generic aspect of our modeling strategy could be exploited for other climate studies that focus on differencing smooth trends and abrupt discontinuities. We discussed the adequation of our method for homogenization problem; of course, our assumption that change points have to occur simultaneously in time could be a limitation in homogenization. Future research is needed to modify our algorithm in order to tailor it to a specific homogenization case study.

Appendix A: Calculations of the Nonlinear Kalman Smoother

[35] For information on the calculations below, we were inspired by different books such as books by *Hossack et al.* [1999] or *Basseville and Nikiforov* [1993].

[36] To simplify the writing, we are using the obvious following notations: $\hat{X}(t|Y_{1:t}) \doteq \mathbb{E}[X_t|Y_{1:t}]$ for the expectation of X_t conditioned on $Y_{1:t}$, and $\hat{\Sigma}(t|Y_{1:t}) \doteq \mathbb{V}ar[X_t|Y_{1:t}]$ for the variance of X_t conditioned on $Y_{1:t}$.

[37] 1st step (prediction step): The first step of the KF solution begins with the estimation of the prediction. It means the calculation of the recurrence relation between $\hat{X}(t|Y_{1:t-1}, b_t)$ and $\hat{X}(t-1|Y_{1:t-1}, b_t)$ and also the similar relation for the variance $\hat{\Sigma}(t|Y_{1:t-1}, b_t)$ and $\hat{\Sigma}(t-1|Y_{1:t-1}, b_t)$ $\forall i = 0,1$

$$\begin{aligned}\hat{X}(t|Y_{1:t-1}, b_t = i) &= \mathbb{E}[\Phi X_{t-1} + E_t|Y_{1:t-1}, b_t = i] \\ &= \Phi\mathbb{E}[X_{t-1}|Y_{1:t-1}, b_t = i] \\ &\quad + \mathbb{E}[E_t|Y_{1:t-1}, b_t = i] \\ &= \Phi\hat{X}(t-1|Y_{1:t-1}) + W_i\end{aligned}$$

and

$$\hat{\Sigma}(t|Y_{1:t-1}, b_t = i) = \Phi\hat{\Sigma}(t-1|Y_{1:t-1})\Phi' + \mathbb{C}ov(W_i) \tag{A1}$$

where

$$W_0 = [0..0], W_1 = [\mu\ \mu\ 0..0],$$

$$\mathbb{C}ov(W_i) = \begin{bmatrix} \Gamma_i & 0 & 0 & 0 \\ 0 & \mathbb{C}ov(E_{fj}) & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \mathbb{C}ov(E_{fj}) \end{bmatrix},$$

$$\mathbb{C}ov(E_{fj}) = \lambda_j\sigma_j^2\begin{bmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{bmatrix}$$

$$\Gamma_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \text{ and } \Gamma_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

[38] 2nd step: Here is the calculation of the predicted distribution of the observations at time t conditioned on the available observations at this time and the occurrence of a change point.
$\forall i = 0,1$

$$\begin{aligned}\mathbb{E}[Y_t|Y_{1:t}, b_t = i] &= \mathbb{E}[HX_t + E_t|Y_{1:t-1}, b_t = i] \\ &= H\hat{X}(t|Y_{1:t}, b_t = i) \\ \mathbb{V}ar[Y_t|Y_{1:t-1}, b_t = i] &= H\hat{\Sigma}(t|Y_{1:t-1}, b_t = i)H' \\ &\quad + \mathbb{V}ar(E_t)\end{aligned} \tag{A2}$$

We can estimate distribution at time t with the same distribution but also conditioned on the occurrence of a break by the approximate mixture of multivariate normal distribution:

$$\begin{aligned}&(Y_t|Y_{1:t}) \text{ is equal in distribution to} \\ &(1-\pi)(Y_t|Y_{1:t}, b_t = 0) + \pi(Y_t|Y_{1:t}, b_t = 1)\end{aligned} \tag{A3}$$

[39] 3rd step: Calculation of the probability q_t^i: This calculation provides the posterior probability at every time of the occurrence of a change point conditioned on the observations available at this same time $Y_{1:t}$

$$\begin{aligned}q_t^0 &\doteq Pr(b_t = 0|Y_{1:t}) = \frac{(1-\pi)Pr(Y_t|Y_{1:t-1}, b_t = 0)}{Pr(Y_t|Y_{1:t-1})} \\ q_t^1 &\doteq Pr(b_t = 1|Y_{1:t}) = \frac{\pi Pr(Y_t|Y_{1:t-1}, b_t = 1)}{Pr(Y_t|Y_{1:t-1})}\end{aligned} \tag{A4}$$

[40] 4th step (update state): This step is the second main of the KF. It deals with the update of the dynamics with the observations of the current time. We have to express a relation between $\hat{X}(t|Y_{1:t}, b_t = i)$ and $\hat{X}(t|Y_{1:t-1}, b_t = i)$ and a similar relation for the second order: $\hat{\Sigma}(t|Y_{1:t}, b_t = i)$ and $\hat{\Sigma}(t|Y_{1:t-1}, b_t = i)$
$\forall i = 0,1$

$$\begin{aligned}\hat{X}(t|Y_{1:t}, b_t = i) &= \hat{X}(t|Y_{1:t-1}, b_t = i) \\ &\quad + \hat{\Sigma}(t|Y_{1:t-1}, b_t = i)H'\mathbb{V}ar([Y_t|Y_{1:t-1}, b_t = i)^{-1} \\ &\quad \cdot [Y_t - \mathbb{E}[Y_t|Y_{1:t-1}, b_t = i]] \\ \hat{\Sigma}(t|Y_{1:t}, b_t = i) &= \hat{\Sigma}(t|Y_{1:t-1}, b_t = i) \\ &\quad - \hat{\Sigma}(t|Y_{1:t-1}, b_t = i)H'\mathbb{V}ar([Y_t|Y_{1:t}, b_t = i)^{-1}H\hat{\Sigma}(t|Y_{1:t-1}, b_t = i)\end{aligned} \tag{A5}$$

[41] 5th step: Finally, we introduce the probability q_t^i to take into account the nonlinearities of the occurrence of a change point at time t. This second order of this step is achieved by using some well-known results on conditional variance, described by *Hossack et al.* [1999].

$$\begin{aligned}\hat{X}(t|Y_{1:t}) &= q_t^0\hat{X}(t|Y_{1:t}, b_t = 0) + q_t^1\hat{X}(t|Y_{1:t}, b_t = 1) \\ \hat{\Sigma}(t|Y_{1:t}) &= \sum_{i=0}^{1}\hat{\Sigma}(\mathbb{E}[X|Y_{1:t}, b_t = i]) + \mathbb{E}\big(\hat{\Sigma}[X|Y_{1:t}, b_t = i]\big) \\ &= \sum_{i=0}^{1} q_t^i\hat{\Sigma}(t|Y_{1:t}, b_t = i) \\ &\quad + q_t^i\big[\hat{X}(t|Y_{1:t}, b_t = i) - \hat{X}(t|Y_{1:t})\big]^2\end{aligned} \tag{A6}$$

[42] 6th step: the Fixed Interval Smoother: Also called the Kalman filtering, this method permits to reconstruct the different components of the state vector given the entire time series, i.e., OLR or ITD). for all t from 1 to T, the FIS constructs $(X_t|Y_{1:T})$. We can obtain the equation:

$$\begin{aligned}\hat{X}(t|Y_{1:T}) &= \hat{X}(t|Y_{1:t}) + C_t\big[\hat{X}(t+1|Y_{1:T}) - \hat{X}(t+1|Y_{1:t})\big] \\ \hat{\Sigma}(t|Y_{1:T}) &= \hat{\Sigma}(t|Y_{1:t}) + C_t\big[\hat{\Sigma}(t+1|Y_{1:T}) - \hat{\Sigma}(t+1|Y_{1:t})\big]C_t'\end{aligned} \tag{A7}$$

where

$$C_t = \hat{\Sigma}(t|Y_{1:t})\Phi\hat{\Sigma}(t+1|Y_{1:t})^{-1} \tag{A8}$$

The algorithm underlying this method was made with the free software environment for statistical computing and graphics "R." "R" provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. For more details, see *R Development Core Team* [2009].

[43] **Acknowledgments.** The authors acknowledge the support of the Laboratoire Atmosphères, Milieux, Observations Spatiales (LATMOS, http://www.latmos.ipsl.fr), a geosciences laboratory belonging to the CNRS/IPSL, and also the support of the GEOmon project (http://www.geomon.eu). The programs used were made with the functional language and environment R (http://www.r-project.org) [see *R Development Core Team*, 2009]. Part of this work has been supported by the EU-FP7 ACQWA Project (http://www.acqwa.ch) under contract 212250, by the PEPER-GIS project, and by the ANR-MOPERA project.

References

Basseville, M., and I. V. Nikiforov (1993), *Detection of Abrupt Changes: Theory and Application*, Prentice-Hall, Englewood Cliffs, N. J., ISBN:0-13-126780-9.

Beaulieu, C., T. B. M. J. Ouarda, and O. Seidou (2007), Synthese des techniques d'homogeneisation des series climatiques, *Hydrol. Sci. J.*, *52*, 18–37, doi: 10.1623/hysj.52.1.18.

Caussinus, H., and O. Mestre (2004), Detection and correction of artificial shifts in climate series, *J. R. Stat. Soc.*, *53*, 405–425, doi: 10.1111/j.1467-9876.2004.05155.x.

Chib, S. (1998), Estimation and comparison of multiple change-point models, *J. Econometr.*, *86*, 221–241, doi: 10.1016/S0304-4076(97)00115-2.

Davis, R. A., T. C. Lee, and G. A. Rodriguez-Yam (2006), Structural break estimation for non-stationary time series signals, *J. Am. Stat. Assoc.*, *101*, 229–239.

Evensen, G. (2006), *Data Assimilation: The Ensemble Kalman Filter*, Springer, New York, ISBN 354038300X.

Fontaine, B., and S. Louvet (2006), Sudan-Sahel rainfall onset: Definition of an objective index, types of years, and experimental hindcasts, *J. Geophys. Res.*, *111*, D20103, doi:10.1029/2005JD007019.

Fontaine, B., S. Louvet, and P. Roucou (2008), Definition and predictability of an OLR-based West African monsoon onset, *Int. J. Climatol.*, *28*, 1787–1798.

Guo, W., Y. Wang, and M. B. Brown (1998), A signal extraction approach to modeling hormone time series with pulses and a changing baseline, *J. Am. Stat. Assoc.*, *94*, 746–756.

Hannart, A., and P. Naveau (2009), Bayesian multiple change points and segmentation: Application to homogenization of climatic series, *Water Resour. Res.*, *45*, W10444, doi: 10.1029/2008WR007689.

Hossack, I. B., J. H. Pollard, and B. Zehnwirth (1999), *Introductory Statistics With Applications in General Insurance*, Cambridge Univ. Press, Cambridge, U. K., ISBN:0 521 65534 X.

Kalman, R. E. (1960), A new approach to linear filtering and prediction problems, *J. Basic Eng.*, *82*, 35–45.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S.-K. Yang, J. Hnilo, M. Fiorino, and G. Potter (2002), NCEP-DOE AMIP-II reanalysis (R-2), *Bull. Am. Meteorol. Soc.*, *83*, 1631–1643.

Lavielle, M., and E. LeBarbier (2001), An application of MCMC methods for the multiple change-points problem, *Signal Proc.*, *81*, 39–53, doi:10.1016/S0165-1684(00)00189-4.

Le Barbé, L., T. Lebel, and D. Tapsoba (2002), Rainfall variability in West Africa during the years 1950–90, *J. Clim.*, *15*(2), 187–202, doi: 10.1175/1520-0442(2002)0152.0.CO;2.

Liebmann, B., and C. Smith (1996), Description of a complete outgoing longwave radiation data set, *Bull. Am. Meteorol. Soc.*, *77*, 1275–1277.

Meinhold, R. J., and N. D. Singpurwalla (1983), Understanding the Kalman Filter, *Am. Statistician*, *37*, 123–127.

Nicholson, S. (1981), Rainfall and atmospheric circulation during drought periods and wetter years in West Africa, *Mon. Weather Rev.*, *109*, 2191–2208.

Parzen, E. (1962), On estimation of a probability density function and mode, *Ann. Math. Stat.*, *33*, 1065–1076.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna, ISBN:3-900051-07-0. (Available at http://www.R-project.org)

Reeves, J., J. Chen, X. L. Wang, R. Lund, and Q. Q. Lu (2007), A review and comparison of change-point detection techniques for climate data, *J. Appl. Meteorol. Climatol.*, *46*, 900–915, doi: 10.1175/JAM2493.1.

Sultan, B., and S. Janicot (2000), Abrupt shift of the ITCZ over West Africa and intra-seasonal variability, *Geophys. Res. Lett.*, *27*, 3353–3356.

Sultan, B., and S. Janicot (2003), The West African monsoon dynamics. part II: The preonset and onset of the summer monsoon, *J. Clim.*, *16*(21), 3407–3427, doi: 10.1175/1520-0442(2003)0162.0.CO;2.

Thorncroft, C., and K. Hodges (2001), African easterly wave variability and its relationship to Atlantic tropical cyclone activity, *J. Clim.*, *14*(6), 1166–1179.

Wahba, G. (1978), Improper priors, spline smoothing and the problem of guarding against model errors in regression, *J. R. Stat. Soc.*, *40*, 364.

Wecker, W. E., and C. F. Ansley (1983), The signal extraction approach to nonlinear regression and spline smoothing, *J. Am. Stat. Ass.*, *78*, 81–89.

Welch, G., and G. Bishop (1995), An introduction to the Kalman Filter, technical report, Univ. of N. C., Chapel Hill, N. C.

E. Flaounas and J. Gazeaux, LATMOS, IPSL, UPMC, UVSQ, CNRS/INSU, Paris F-75005, France. (julien.gazeaux@latmos.ipsl.fr)

A. Hannart, IFAECI, Universidad de Buenos Aires, CNRS/CONICET, 1428 Buenos Aires, Argentina.

P. Naveau, LSCE, IPSL, CNRS/CEA, Gif-Sur-Yvette F-91191, France.